eDNA 监测测序数据分析注释中参考数据库选择、指标阈值选择、目标数据准备的影响——以长江中游鱼类为监测目标*

许兰馨[1,2]，杨海乐[1]**，刘志刚[1]，杜　浩[1]

（1.中国水产科学研究院长江水产研究所，农业农村部淡水生物多样性保护重点实验室，湖北　武汉 430223

2.南京农业大学无锡渔业学院，江苏　无锡　214000）

摘　要：在基于宏条形码（meta-barcoding）的 eDNA 监测技术路径中，eDNA 测序数据的分析和注释是决定监测结果判断和评估精确与否甚至准确与否的基础，而参考数据库选择、指标阈值选择、目标数据准备是 eDNA 测序数据分析和注释中最为关键的 3 个技术环节。为弄清上述 3 个技术环节处理方案的影响，本研究以长江中游 2 组 eDNA 监测 *COI* 基因测序数据为分析对象，针对鱼类的检出做了 3 组实验来分别检验 1）不同参考数据库及物种注释算法对注释结果的影响，2）不同 OTU 聚类序列相似度和物种注释分类置信度（序列一致性和序列覆盖度）对注释结果的影响，3）目标数据中各物种不同序列丰富度对注释结果的影响。结果显示：1）Blast 算法下，3 个版本 nt 库注释出的物种基本一致（72%~78%），2 个本地序列参考库注释出的物种也基本一致（91%~96%），这 5 个序列参考库注释出的物种 52%~68%一致；nt 库 RDP Classifier 算法注释出的物种覆盖 95%以上 Blast 算法注释出的物种，并比 Blast 算法注释出的物种多 151%~443%，多出的物种大都是错误注释，本地参考数据库 RDP Classifier 算法注释出的物种覆盖 66%~85%的 Blast 算法注释出的物种，并存在数条只注释到科属的结果。2）OTU 聚类序列相似度阈值，取值 0.999 比取值 0.99 获得的 OTU 多 154%~209%，注释到鱼类的 OTU 多 240%~490%；注释分类置信度阈值（Blast 算法，序列一致性和序列覆盖度）从 0.8 到 0.99 注释获得的物种组成基本（94%以上）一致，物种下的 OTU 组成也基本（83%以上）一致，注释分类置信度阈值取 0.7 时注释获得的物种组成、OTU 组成和取 0.8 及以上时注释获得的有较大差异。3）在 OTU 聚类序列相似度阈值 0.999、注释分类置信度阈值 0.9 时，多序列数据注释所得鱼类物种数、OTU 数最多、物种注释正确率最高（达 81.49%），分别比单序列数据的多 7%、215%和高 5%。在具体 eDNA 测序数据的分析和注释中，可通过建立完善本地参考数据库、优化 OTU 聚类序列相似度和物种注释分类置信度（序列一致性和序列覆盖度）取值、增加目标数据的丰富度来提高注释结果的准确性，但受制于物种注释算法的局限性，物种注释错误和注释遗漏的问题可能将长期存在，物种注释正确率通常低于 85%（基于 *COI* 基因的 eDNA 监测）。

关键词：环境 DNA，鱼类，宏条形码，参考数据库，OTU 聚类序列相似度，物种注释分类置信度，长江中游

The impacts of reference database selection, indicator threshold determination and target data preparation in the sequence data analysis of eDNA monitoring -- taking fish as the target in middle Yangtze River

Xu Lanxin [1,2], Yang Haile [1] **, Liu Zhigang [1] & Du Hao [1]

(1. *Key Laboratory of Freshwater Biodiversity Conservation, Ministry of Agriculture and Rural Affairs, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan* 430223, *P.R.China*

2. *Wuxi Fisheries College, Nanjing Agricultural University, Wuxi* 214000, *P.R.China*)

---

* 中央级公益性科研院所基本科研业务费专项（YFI202201）、农业农村部财政专项“长江禁捕后常态化监测”资助。

** 通信作者；E-mail：haileyang10@fudan.edu.cn。

Abstract: In the meta-barcoding based eDNA monitoring technology, the analysis and annotation of eDNA sequencing data serve as the foundation for obtaining accurate and reliable monitoring results. The selection of reference databases, the determination of analysis & annotation indicator thresholds, and the preparation of target data are the most critical technical steps in eDNA sequencing data analysis and annotation. To clarify the impacts of these three technical aspects and provide scientific support for the standardization of eDNA monitoring technology, the current study used two sets of *COI* gene sequence data from eDNA monitoring in the middle reach of the Yangtze River as the analysis objects and designed three sets of experiments to test 1) the impacts of different reference databases and species annotation algorithms on the annotation results, 2) the impacts of different OTU clustering sequence similarity and species annotation classification confidence (sequence consistency and sequence coverage) on the annotation results, and 3) the impacts of different target sequence data richness of each species on the annotation results. The results showed that: 1) under the Blast algorithm, the annotated species matched with three versions of nt library from NCBI were generally consistent (72%~78%); those matched with two local sequence reference libraries were also generally consistent (91%~96%); and the annotated species from the five result matched with these five sequence reference libraries were consistent in 52%~68%. The RDP Classifier algorithm annotated species matched with nt libraries covered over 95% of Blast algorithm annotated species, and increased by 151%~443% species, but most additional species were misannotated. The RDP Classifier algorithm annotated species matched with local sequence reference libraries covered 66%~85% of Blast algorithm annotated species, and there were several results only annotated to family or genus level. 2) When the OTU clustering sequence similarity threshold was set to 0.999, it obtained 154%~209% more OTUs than when set to 0.99, and 240%~490% more annotated OTUs of fish were obtained. The classification confidence threshold (Blast algorithm) had little effect on species composition when changed from 0.8 to 0.99, with over 94% consistency, but there was a significant difference when it was set to 0.7. 3) When the OTU clustering sequence similarity threshold was 0.999 and the classification confidence threshold was 0.9, the number of fish species and OTUs obtained from multiple sequences data annotation was the largest, and had the highest species annotation accuracy (81.49%), which increased by 7% fish species, 215% OTUs and 5% accuracy respectively compared to single sequences data annotation. In eDNA sequencing data analysis and annotation, accuracy can be improved by establishing and improving local reference databases, optimizing OTU clustering sequence similarity and species annotation classification confidence thresholds (sequence consistency and sequence coverage), increasing target sequence data richness. However, due to the limitation of species annotation algorithms, problems such as species annotation errors and omissions may persist in eDNA sequencing data analysis and annotation in the future. Then, the species annotation accuracy of eDNA monitoring (based on the *COI* gene) would always lower than 85%.

Keywords: environmental DNA; fish; meta-barcoding; reference database; OTU clustering sequence similarity; species annotation classification confidence; middle Yangtze River

eDNA (environmental DNA)是指从环境样品（水体、土壤、沉积物、空气、混合物等）中提取的 DNA，是各种生物的 DNA 混合物[1-3]。从环境样品中提取 eDNA，用特定 DNA 宏条形码(metabarcoding)引物对其进行扩增测序、分类学分析、相对丰度分析、功能预测等，可以监测环境中物种组成、群落结构、生态功能等相关信息[3-5]。近年来随着宏条形码技术的成熟、二代测序技术成本的下降[3]，eDNA 监测工作有向常态化发展的趋势[6]，其中在禁捕水域针对鱼类物种组成及资源开展 eDNA 监测的需求最为迫切。

实现 eDNA 监测工作常态化的前提是实现 eDNA 监测的标准化[7, 8]。eDNA 监测技术链条中，eDNA 测序结果的分析注释作为监测工作的出口控制环节，是整个监测工作后续结果判断和评估的基础[3]，决定了监测结果判断和评估的精确度和准确度。参考数据库选择、阈值指标选择、目标数据准备是 eDNA 测序数据的分析和注释中最为关键的 3 个技术环节，其重要性和影响已有整体性论述[3]，截至目前的案例研究

也都有自身选择（参考数据库以 NCBI 的 nt 数据库为主[9-13]，小部分进行自建本地参考数据库[14]；OTU 聚类的序列相似度以 0.97、0.99 为主[12, 14]，也有用 0.95、0.98、1.00 的[10, 13]；物种注释中的序列覆盖度取值有 0.80、0.85、0.95，序列一致性取值有 0.95、0.96、0.97、0.99、1.00[10, 12, 14]；也有不少研究不标明相关技术环节参数[9, 11]；我们随机抽取了一些案例研究的相关参数信息于附表 1 中），比较多样化，结果可信度、可比性存疑，亟需标准化。但若要满足具体的标准化，尚需进行具体定量比较分析。

针对这 3 个技术环节的标准化需求，本研究以长江中游 2 组 eDNA 监测 *COI* 基因序列数据为分析对象，针对鱼类的检出做了 3 组实验来分别检验 1）不同参考数据库及物种注释算法对注释结果的影响，2）OTU 聚类序列相似度和物种注释分类置信度（序列一致性和序列覆盖度）对注释结果的影响，3）目标数据中各物种序列丰富度对注释结果的影响，并探讨其中的影响机制。

1 材料方法

1.1 两个数据集的来源

2020 年 6 月在长江中游段设置 30 个采样断面，相邻采样断面间的径流距离为 30 km 左右，乘船在江中间按照采样断面用无菌采样瓶采取上层水水样 1.5 L。在正式采集水样并封装之前，用所要采的水涮洗 3 次。水样在冰浴中暂存，在实验室中用 0.2 μm 孔径滤膜进行抽滤（采样后 6 h 内处理完毕），获得存留了 eDNA 的滤膜，放入 50 mL 无菌离心管，自封袋装好，﹣80℃冰箱保存，泡沫箱干冰浴运输。委托上海美吉生物医药科技有限公司用线粒体 *COI* 基因的扩增子（引物为 mlCOIintF/jgHCO2198R，片段长度 320 bp 左右）在 Illumina Miseq 平台以双向测序方式进行二代高通量测序，并进行双端序列拼接，获得长江中游段 eDNA 数据集[15]。2020 年 9 月在长江武汉段的 1 个采样断面连续 13 天采集 eDNA 样品并送样测序，具体采样、测序方法同上，获得长江武汉段 eDNA 数据集[16]。相关序列原始数据已存于国家基因库生命大数据平台（China National GeneBank DataBase, CNGBdb, https://db.cngb.org/）的长江中游 eDNA 序列文件夹中（项目编号: CNP0002410，DOI: 10.26036/CNP0002410）。

1.2 本地参考数据库的构建

根据“长江渔业资源与环境调查（2017-2021）”所整理出的长江鱼类名录[17]，在 NCBI 数据库中搜集各物种的线粒体 *COI* 基因序列，并基于 2021 年在长江渔业资源与环境调查中所捕捞采集的各种鱼类的鳍条样品，通过 DNA 提取、用线粒体 *COI* 基因的宏条形码引物 mlCOIintF/jgHCO2198R 进行 PCR 扩增、送武汉天一辉远生物科技有限公司进行序列测定，获得相关鱼类物种的线粒体 *COI* 序列，整合构建本地针对长江鱼类的线粒体 *COI* 基因的宏条形码引物 mlCOIintF/jgHCO2198R 参考数据库。共收集到隶属于 18 目 41 科 149 属 281 种的 2040 条线粒体 *COI* 序列（附表 2），其中从 NCBI 搜集获得 236 个物种共 1744 条序列（截至 2022 年 3 月），自行扩增、测序获得 115 个物种共 296 条序列（【金山文档】 长江鱼类 COI 基因条形码-持续更新 https://kdocs.cn/l/cgi2zpUmWHaS）。用所收集到的所有参考序列（共 281 种 2040 条）构建本地多序列参考库，从本地多序列参考库中对每一个物种随机抽取一条参考序列（共 281 种 281 条）构建本地单序列参考库。

1.3 不同参考数据库及物种注释算法的注释结果比较分析

利用美吉生物云平台（www.majorbio.com）的分析计算模块，对长江中游段 eDNA 监测 *COI* 数据、长江武汉段 eDNA 监测 *COI* 数据进行质控、拼接（用 FLASH version 1.2.11 https://ccb.jhu.edu/software/FLASH/index.shtml）、OTU 聚类（用 USEARCH7-uparse http://drive5.com/uparse/，取 99% 的序列相似度）、物种注释（用 Blast 算法，RDP Classifier 算法 version 2.11 https://sourceforge.net/projects/rdp-classifier/）。以 NCBI 核酸序列数据库 nt_v20200604、nt_v20210917 库、

nt_v20221012 库、本地多序列参考库、本地单序列参考库为参考数据库，采用 Blast 算法和 RDP Classifier 算法取 90%的分类置信度（Blast 算法中为序列一致性和序列覆盖度阈值均取 90%）进行物种注释，获得注释结果，筛选出辐鳍鱼纲（Actinopteri）结果，最后进行基于不同参考数据库所获得注释结果（鱼类物种数、鱼类物种组成、鱼类 OTU 数）的比较分析。由于美吉生物云平台 2022 年更新，对 nt 库不再提供 RDP Classifier 算法注释的分析接口，故未对基于 nt_v20221012 库的物种注释进行 RDP Classifier 算法注释的分析。

1.4 不同 OTU 聚类序列相似度和物种注释分类置信度的注释结果比较分析

基于 OTU 聚类序列相似度和物种注释分类置信度的取值规则及常用取值，对 OTU 聚类序列相似度和物种注释分类置信度（用 Blast 算法，即序列一致性和序列覆盖度，两者取同一个值）取值设定 14 个组合，unoise3 & 0.99、unoise3 & 0.9、unoise3 & 0.8、unoise3 & 0.7、0.999 & 0.99、0.999 & 0.9、0.999 & 0.8、0.999 & 0.7、0.99 & 0.97、0.99 & 0.9、0.99 & 0.8、0.99 & 0.7、0.97 & 0.8、0.9 & 0.8，对长江中游段 eDNA 监测 *COI* 数据、长江武汉段 eDNA 监测 *COI* 数据进行分析注释，参考数据库用 NCBI 核酸序列数据库的 nt_v20221012 库，获得注释结果，筛选出辐鳍鱼纲（Actinopteri）结果，进行不同阈值取值所获得注释结果（鱼类物种数、鱼类物种组成、鱼类 OTU 数、获得物种注释的 OTU 比例）的比较分析。分析计算平台、模块、步骤见 1.3。

1.5 目标数据中各物种不同序列丰富度的注释结果比较分析

基于本地单序列参考库、本地多序列参考库分别构建 2 个目标数据。考虑到在分析计算步骤中的 OTU 聚类环节默认去除无重复序列，所以在构建目标数据过程中对参考序列进行 7 倍重复。OTU 聚类序列相似度和物种注释分类置信度（用 Blast 算法，即序列一致性和序列覆盖度，两者取同一个值）取值设定 4 个组合，0.999 & 0.99、0.999 & 0.9、0.99 & 0.9、0.99 & 0.8，对 2 个目标数据进行物种注释，参考数据库用 NCBI 核酸序列数据库的 nt_v20221012 库，获得注释结果，筛选出辐鳍鱼纲(Actinopteri)结果，进行不同目标数据所获得注释结果（鱼类物种数、鱼类物种组成、鱼类 OTU 数、物种正确注释比例）的比较分析。分析计算平台、模块、步骤见 1.3。

2 结果

2.1 参考数据库及物种注释算法对注释结果的影响

针对长江中游段 eDNA 监测 *COI* 数据、长江武汉段 eDNA 监测 *COI* 数据这 2 个数据集进行的 3 个 nt 参考数据库 Blast 算法注释结果比较分析显示，3 个 nt 参考数据库注释到的鱼类物种数、鱼类物种组成、鱼类 OTU 数差异不大（图 1，附图 1a，b）。长江中游段注释到的鱼类物种数为 27~29、鱼类 OTU 数为 47~48（图 1），共涉及鱼类物种 35 个，其中共有的物种 21 个，注释结果的物种组成一致性可达 72%（附图 1a）。长江武汉段注释到的鱼类物种数为 32~34、鱼类 OTU 数为 95（图 1），共涉及鱼类物种 41 个，其中共有的物种 25 个，注释结果的物种组成一致性可达 78%（附图 1b）。

针对 2 个数据集进行的 2 个本地参考数据库 Blast 算法注释结果比较分析显示，多序列参考库和单序列参考库注释到的鱼类物种数、鱼类物种组成、鱼类 OTU 数基本一致（图 1，附图 1a，b）。长江中游段注释到的鱼类物种数为 25、鱼类 OTU 数为 44（图 1），共涉及鱼类物种 26 个，其中共有的物种 24 个，注释结果的物种组成一致性可达 96%（附图 1a）。长江武汉段注释到的鱼类物种数为 31~32、鱼类 OTU 数为 93~94（图 1），共涉及鱼类物种 34 个，其中共有的物种 29 个，注释结果的物种组成一致性可达 91%（附图 1b）。

比较分析 2 个数据集的 3 个 nt 参考数据库和 2 个本地参考数据库 Blast 算法注释结果，nt 参考数据库

和本地参考数据库注释到的鱼类物种数、鱼类 OTU 数差异不大（图 1），但鱼类物种组成存在系统性差异（附图 1a，b）。长江中游段用 3 个 nt 参考数据库和 2 个本地参考数据库到的鱼类共涉及 42 种，其中共有鱼类物种数为 15，注释结果的物种组成一致性约 52%（附图 1a）。长江中游段用 3 个 nt 参考数据库和 2 个本地参考数据库到的鱼类共涉及 51 种，其中共有鱼类物种数为 21，注释结果的物种组成一致性约 68%（附图 1b）。

2 个数据集的 2 个 nt 参考数据库 RDP Classifier 算法注释结果，鱼类物种组成基本覆盖 3 个 nt 参考数据库 Blast 算法注释出的鱼类物种，比 Blast 算法注释出的鱼类物种有大量增加（151%~343%，相应的 OTU 也有大量增加），多出的物种大都（99%）不是长江鱼类物种（图 1，附图 1a，b）。长江中游段 RDP Classifier 算法注释到的鱼类物种数为 134~135、鱼类 OTU 数为 189（图 1），共涉及鱼类物种 155 个，覆盖 Blast 算法注释出的所有 21 个共有物种和 35 个所涉物种中的 34 个，覆盖率可达 97%（附图 1a）。长江武汉段 RDP Classifier 算法注释到的鱼类物种数为 88~89、鱼类 OTU 数为 159~165（图 1），共涉及鱼类物种 103 个，覆盖 Blast 算法注释出的所有 25 个共有物种和 41 个所涉物种中的 39 个，覆盖率可达 95%（附图 1b）。

2 个数据集的 2 个本地参考数据库 RDP Classifier 算法注释结果，鱼类物种组成覆盖大部分 2 个本地参考数据库 Blast 算法注释出的鱼类物种，比 Blast 算法注释出的鱼类物种有所减少，尤其是单序列本地参考库注释出的鱼类物种数明显减少，OTU 数有所增加，所增加 OTU 多为只注释到科属的结果（s__unclassified_）（图 1，附图 1a，b）。长江中游段 RDP Classifier 算法注释到的鱼类物种数为 21~25、鱼类 OTU 数为 46~192（图 1），共涉及鱼类物种 28 个，其中共有物种 18 个，覆盖 Blast 算法注释出的 24 个共有物种中的 16 个和 26 个所涉物种中的 22 个，覆盖率约 67%~85%（附图 1a）。长江武汉段 RDP Classifier 算法注释到的鱼类物种数为 24~32、鱼类 OTU 数为 65~291（图 1），共涉及鱼类物种 35 个，其中共有物种 21 个，覆盖 Blast 算法注释出的 29 个共有物种中的 19 个和 34 个所涉物种中的 28 个，覆盖率约 66%~82%（附图 1b）。RDP Classifier 算法下，多序列参考库注释出的鱼类物种数、OTU 数明显比单序列参考库注释出的多，但所多的部分大多与只注释到科属的结果相关（附图 1a，b）。

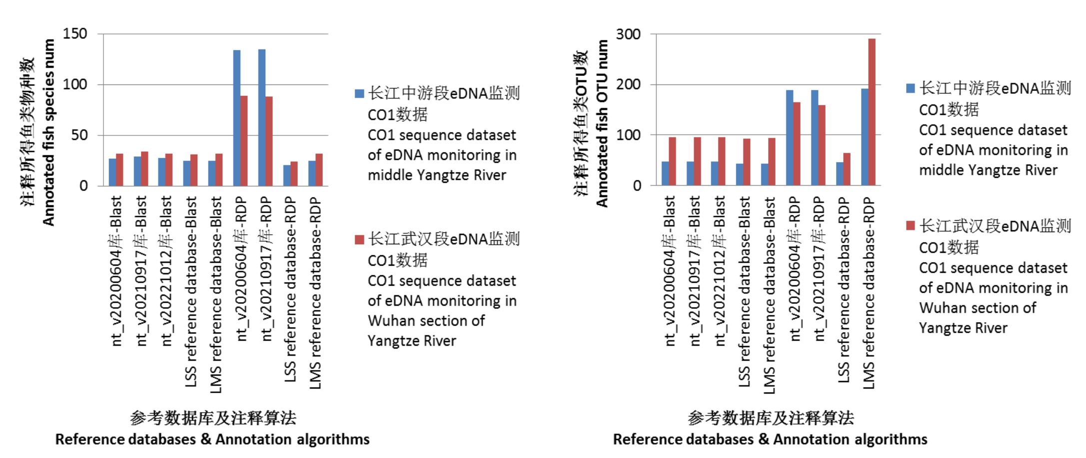


图 1 不同版本 nt 参考数据库、本地参考库及相应注释算法对注释结果的影响

Fig. 1 The influence of different visions of nt reference database from NCBI and of local reference databases and of annotation algorithms on annotation results

注：LSS，本地单序列；LMS，本地多序列

Notes: LSS, local single-sequence; LMS, local multi-sequence

2.2 OTU 聚类序列相似度和物种注释分类置信度对注释结果的影响

对 2 个数据集的不同 OTU 聚类序列相似度和物种注释分类置信度（Blast 算法，序列一致性和序列覆盖度，两者取同一个值）取值组合所获得的鱼类注释结果比较分析显示，OTU 聚类序列相似度阈值越高，所获得的 OTU 数量越多，物种注释分类置信度阈值越高，匹配到参考序列上 OTU 的越少（获得物种注释的 OTU 比例越低），但在物种注释分类置信度 0.8—0.99 区间注释所得鱼类物种数和物种组成基本一致，OTU 数也差异不大，而物种注释分类置信度 0.7 注释所得鱼类物种数、物种组成、OTU 数则和 0.8 及以上的有明显差别（图 2，附图 2a~d）。

在 OTU 聚类分析中，用序列相似度 0.999，0.99，0.97，0.9 对长江中游段 eDNA 监测 *COI* 数据聚类获得 OTU 数量为 45637、14783、10698、6129，对长江武汉段 eDNA 监测 *COI* 数据聚类获得 OTU 数量为 38766、15235、7541、4120。去噪分析所获得的 OTU 数量比序列相似度 0.9 的更少一点，对长江中游段 eDNA 监测 *COI* 数据聚类获得 OTU 数量为 4648，对长江武汉段 eDNA 监测 *COI* 数据聚类获得 OTU 数量为 3573。

在物种注释分析中，分类置信度从 0.7 到 0.8，0.9，0.99，获得物种注释的 OTU 的占比逐步降低，在分类置信度大于 0.9 时，获得物种注释的 OTU 比例基本都在 55%以下。分类置信度在 0.8—0.99 区间注释所得的鱼类物种数差异不大于 6%， OTU 数差异不大于 17%（图 2），物种组成差异不大于 6%（附图 2a，c）。与 0.8—0.99 分类置信度注释结果相比，0.7 的分类置信度注释所得的鱼类物种数多 47%~116%，OTU 数多 23%~98%（图 2），物种组成覆盖所有 0.8—0.99 分类置信度注释到的物种，并比 0.8—0.99 分类置信度注释到的结果多出一系列错误注释的物种（附图 2a，c）。

在 OTU 聚类和物种注释的组合分析中，大于等于 0.97 的 OTU 聚类序列相似度和大于等于 0.8 的物种注释分类置信度的组合能够注释获得比较可靠的鱼类物种组成，组合间的物种组成一致性可达 88%以上（附图 2 a，c）。大于等于 0.99 的 OTU 聚类序列相似度和大于等于 0.8 的物种注释分类置信度的组合能够注释获得可以反映鱼类种内遗传多样性的 OTU 数量（图 2，附图 2a，c）。由 2 个数据集反映的在长江中游监测到具有较高遗传多样性的鱼类有黑尾近红鲌（*Ancherythroculter nigrocauda*）、草鱼（*Ctenopharyngodon idella*）、贝氏䱗（*Hemiculter bleekeri*）、鲢（*Hypophthalmichthys molitrix*）、鳊（*Parabramis pekinensis*）、似鳊（*Pseudobrama simoni*）、寡鳞飘鱼（*Pseudolaubuca engraulis*）、赤眼鳟（*Squaliobarbus curriculus*）、银鲴（*Xenocypris argentea*）等（附图 2a，c）。

在 OTU 聚类和物种注释的组合分析中，0.999 的 OTU 聚类序列相似度和大于等于 0.8 的物种注释分类置信度的组合所注释获得的鱼类序列数，明显比 0.9~0.99 的 OTU 聚类序列相似度和大于等于 0.8 的物种注释分类置信度的组合所注释获得的鱼类序列数要少，总数量少 76%~81%，其中导致差距的主要是个别几个物种的序列差异，比如长江中游段数据集中的贝氏䱗（*Hemiculter bleekeri*）、鲢（*Hypophthalmichthys molitrix*）、寡鳞飘鱼（*Pseudolaubuca engraulis*），长江武汉段数据集中的黑尾近红鲌（*Ancherythroculter nigrocauda*）、草鱼（*Ctenopharyngodon idella*）、银鲴（*Xenocypris argentea*）（附图 2b，d），均属数据中遗传多样性最高的几个物种，其它大多数物种的序列数少 10%~30%。

数据集 Dataset	OTU聚类序列相似度 Similarity	物种注释分类置信度 Confidence	注释所得鱼类物种数 Fish species num	注释所得鱼类OTU数 Fish OTU num	获得物种注释的OTU比例 Percent of annotated OTU
长江中游段eDNA监测CO1数据 CO1 sequence dataset of eDNA monitoring in middle Yangtze River	去噪 Unoise	0.99	17	20	10.82%
	去噪 Unoise	0.9	17	20	35.50%
	去噪 Unoise	0.8	17	20	66.01%
	去噪 Unoise	0.7	24	27	81.07%
	0.999	0.99	28	252	26.27%
	0.999	0.9	29	283	53.95%
	0.999	0.8	29	283	71.45%
	0.999	0.7	54	311	83.24%
	0.99	0.97	28	45	15.13%
	0.99	0.9	28	48	27.82%
	0.99	0.8	28	48	56.56%
	0.99	0.7	54	75	75.01%
	0.97	0.8	28	28	50.69%
	0.9	0.8	19	19	35.24%
长江武汉段eDNA监测CO1数据 CO1 sequence dataset of eDNA monitoring in Wuhan section of Yangtze River	去噪 Unoise	0.99	22	25	4.81%
	去噪 Unoise	0.9	23	27	18.25%
	去噪 Unoise	0.8	23	27	49.51%
	去噪 Unoise	0.7	26	31	68.99%
	0.999	0.99	31	303	6.84%
	0.999	0.9	33	363	28.74%
	0.999	0.8	33	364	52.54%
	0.999	0.7	67	599	69.40%
	0.99	0.97	31	89	11.27%
	0.99	0.9	32	95	22.35%
	0.99	0.8	32	95	46.13%
	0.99	0.7	51	152	64.40%
	0.97	0.8	31	37	42.71%
	0.9	0.8	16	16	30.61%

图 2 聚类序列相似度和物种注释分类置信度对注释结果的影响

Fig. 2 The influence of different OTU cluster sequence similarity and different species annotation classification confidence on annotation results

注：物种注释分类置信度，在 Blast 算法中为序列一致性和序列覆盖度，两者取同一个值。

Notes: In the Blast algorithm, the classification confidence of species annotation refers to the sequence similarity and sequence coverage, both of them take the same value.

2.3 目标数据中各物种序列丰富度对注释结果的影响

基于本地单序列参考库、本地多序列参考库所构建的 2 个目标数据开展的 OTU 聚类和物种注释分析对比结果显示，在不同分析注释参数条件下均呈现出，目标数据中各物种的序列丰富度越高，注释所得的物种数越多、OTU 数越多，物种正确注释比例越高（图 3）。同时，结果也验证了在对不同物种丰富度下的目标数据注释过程中，OTU 聚类序列相似度越高，所获得的 OTU 越精细，OTU 数量也越多，注释所得

的物种也越多，物种正确注释比例越高（图 3）。

在对 2 个不同物种丰富度的目标数据注释过程中，0.999 的 OTU 聚类序列相似度、0.9 的物种注释分类置信度（Blast 算法，序列一致性和序列覆盖度，两者取同一个值）均获得了最高的注释所得鱼类物种数、注释所得鱼类 OTU 数、物种正确注释比例（图 3）。在最优分析注释参数（0.999&0.9）条件下，多序列数据比单序列数据注释所得鱼类物种数与 OTU 数分别多 7%和 215%，物种正确注释比例高约 4%（图 3）。

单序列数据用 4 组参数注释到的鱼类物种（4 组结果）共涉及 258 种，其中一致的有 237 种；多序列数据用 4 组参数注释到的鱼类物种（4 组结果）共涉及 296 种，其中一致的有 235 种；单序列数据和多序列数据用 4 组参数注释到的鱼类物种（8 组结果）共涉及 299 种，其中一致的有 216 种（附图 3a）。高 OTU 聚类序列相似度和物种注释分类置信度的结果并不完全对低 OTU 聚类序列相似度和物种注释分类置信度的结果构成全覆盖，多序列数据注释所得结果也不完全对单序列数据注释所得结果构成全覆盖（附图 3a）。

单序列数据用 4 组参数一共正确注释到鱼类 220 种，其中 4 组参数下均正确注释到的有 201 种；多序列数据用 4 组参数一共正确注释到鱼类 235 种，其中 4 组参数下均正确注释到的有 206 种；单序列数据和多序列数据用 4 组参数一共正确注释到鱼类 236 种，其中 2 组数据 4 组参数下均正确注释到的有 189 种（附图 3b）。单序列数据和多序列数据用 4 组参数注释到的鱼类（8 组结果）中有 15%~17%的鱼类物种是错误的（附图 3a，b）。单序列数据和多序列数据都未被正确注释的鱼类物种均为本研究扩增获得但 nt 数据库中没有参考序列的物种（附表 2，附图 3b）。

数据集 Dataset	OTU聚类序列相似度 Similarity	物种注释分类置信度 Confidence	注释所得鱼类物种数 Fish species num	注释所得鱼类OTU数 Fish OTU num	物种正确注释比例 Annotation accuracy
本地单序列数据（281个物种，281条序列） Local single-sequence data (281 species, 281 sequences)	0.999	0.99	255	264	77.22%
	0.999	0.9	257	271	77.94%
	0.99	0.9	240	248	72.24%
	0.99	0.8	240	248	72.60%
本地多序列数据（281 个物种，2040条序列） Local multi-sequence data (281 species, 2040 sequences)	0.999	0.99	273	714	80.78%
	0.999	0.9	275	853	81.49%
	0.99	0.9	257	444	77.22%
	0.99	0.8	253	445	76.51%

图 3 目标数据中各物种序列丰富度对注释结果的影响

Fig. 3 The influence of different sequence richness of each species in target sequence data on annotation results

注：物种注释分类置信度，在 Blast 算法中的序列一致性和序列覆盖度，两者取同一个值

Notes: In the Blast algorithm, the classification confidence of species annotation refers to the sequence similarity and sequence coverage, both of them take the same value.

3 讨论

物种注释的 Blast 算法和 RDP Classifier 算法因为计算逻辑不同而获得不同的物种注释结果，但因为指向相同的目标，所以最符合目标的物种注释结果两种算法均可获得，并且 RDP Classifier 算法的结果可以覆盖 Blast 算法的结果，还会多出一些错误匹配的物种。Blast（basic local alignment search tool）算法是基于局部相似性然后拓展延伸到全局的比对算法，并不确保能够找到全局最优解，但所找到的解绝大多数解近乎于是全局最优解，本研究中参考不同版本 nt 库注释获得的结果具有 72%~78%的一致性（附图 1a，b）。RDP Classifier 基于贝叶斯原理进行分类，通过计算后先验概率和条件概率并用 Bootstrap 策略计算其置信

度获得最终注释，结果基本上能够覆盖目标结果，但同时也会获得一些非目标结果的错误注释，本研究中 RDP Classifier 算法 nt 库注释获得的结果可覆盖 95%以上的 Blast 算法注释获得的结果，同时也多 151%~343%的物种，多出的大都（99%）是错误注释的物种（附图 1a，b），类似结果和结论在先前的研究中已有提及和探讨[16]。因此在基于 *COI* 基因的鱼类 eDNA 监测应用中，结果的注释如果选择 Blast 算法，对物种注释结果有必要保持一定的谨慎，物种组成中可能仅有 70%~80%是确定正确的，另外的可能是注释错误，也可能是注释遗漏；如果选择 RDP Classifier 算法，对物种注释结果要保持更大的谨慎，可能大部分物种（60%~80%）都是错误注释的；注释过程中如果参考数据库中缺少相关物种的参考序列，在 Blast 算法下，相关物种的目标序列会归为 s_unclassified，在 RDP Classifier 算法下，相关物种的目标序列会被注释到近源物种或者相关科属下。

在 Blast 算法下，参考数据库对注释结果的影响主要取决于参考数据库内的物种覆盖度和种内变异覆盖度，参考库对目标物种的覆盖越高、每个物种的参考序列越丰富、对物种内序列变异覆盖越全，OTU 比对脱靶的概率越低，能够比对得上的 OTU 数量越多，所得注释结果也越全面，但如果参考库未能有效覆盖相关物种或相关物种的核苷酸序列变异，就容易导致在比对注释过程中部分 OTU 的脱靶、部分物种的注释遗漏。在本研究中，不同版本 nt 库注释获得的物种组成和本地参考数据库注释获得的物种组成在部分物种上具有系统性差异（附图 1a，b），这种系统性差异来自本地参考库和不同版本 nt 库内参考序列物种组成和序列变异覆盖度的系统性差异。长江记录有鱼类 458 种（包含外来种）[17]，其中 222 种鱼类在 nt 库中缺少线粒体 *COI* 基因的扩增子（引物为 mlCOIintF/jgHCO2198R）对应的参考序列，部分有参考序列的物种的序列数量还比较少，对物种内的核苷酸变异覆盖有限（截至 2022 年 3 月），这导致了在实践应用中物种数正确检出的偏少，以及各物种序列相对丰度和实际群落结构的不匹配[16, 18]。为了克服 nt 库中参考序列的不足，可以构建本地参考数据库，快速便捷地更新补充相关物种的参考序列，尽可能全面地覆盖各物种的种内变异。在本地数据库构建方面，南京农业大学等 10 余所高校及科研院所合作构建的中国淡水大型底栖无脊椎动物条形码数据库是类似工作的先行者[19]。建议根据长江鱼类名录[17]，整合 NCBI 的 nt 数据库中已有的相关宏条形码参考序列和本地扩增获得的相关宏条形码参考序列，建立动态补充的本地参考数据库，以尽可能全面地覆盖本地物种及各物种内的核苷酸序列变异。本研究初步构建的长江鱼类条形码本地参考数据库以所有人可访问、可下载、可编辑的在线文档形式作为共享本地数据库供各相关研究者使用，执行 CC-BY-4.0 协议，我们正基于我们已有的长江鱼类标本库持续补充各长江鱼类的参考序列，也欢迎各相关研究者对相关参考序列进行持续补充。

在 Blast 算法下，OTU 聚类序列相似度和物种注释分类置信度取值建议综合考虑物种注释结果的精准度和注释结果的覆盖度。OTU 聚类时序列相似度设置越高，聚类形成的 OTU 就越精细、数目也就越多[3]。序列比对注释的分类置信度设置得偏低，往往会出现把某一物种的序列错误匹配到序列相似的另一物种上；序列比对注释的分类置信度设置得偏高，往往会出现一些通过比对无法与参考数据库中的任一序列形成匹配的 OTU 序列[3]。不同类群物种间的参考序列差异程度有差异，不同物种内的变异程度也有差异，所以在 OTU 聚类和物种注释过程中所适宜的参数设置也会有差异[20]，比如针对细菌的线粒体 16S rRNA 基因的参数取值（比如 0.97 & 0.8 [21]）通常低于针对真核生物的 *COI* 基因的参数取值（比如 0.99 & 0.97 [16]）。本研究针对鱼类（*COI* 基因上的 320 bp 大小的片段）的 2 个数据集分析中，在 OTU 聚类序列相似度大于等于 0.99、物种注释分类置信度（序列一致性和序列覆盖度，两者取同一个值）大于等于 0.8 时，注释所得 OTU 数和物种数的比值大于 1.5（图 2），OTU 能够反映一定的种内核苷酸序列变异（附图 2a，c），注释所得 OTU 数和物种数均相对稳定（图 2），物种注释相对一致（物种组成一致性可达 88%以上，附图 2a，c）。在 OTU 聚类序列相似度为 0.999 时，注释到鱼类的序列数比为 0.99 及以下时少 76%~81%（附图 2b，

d）；在 OTU 聚类序列相似度一定时，注释到鱼类的 OTU 数随着物种注释分类置信度的升高而降低，0.97~0.99 的物种注释分类置信度可获得 6%~26%的 OTU 注释率，0.9 的物种注释分类置信度可获得 22%~54%的 OTU 注释率，0.8 的物种注释分类置信度可获得 46%~71%的 OTU 注释率（图 2）。因此在针对鱼类线粒体 *COI* 基因宏条形码的 eDNA 监测数据分析中，结果的注释如果选择 Blast 算法，建议 OTU 聚类序列相似度在 0.99~0.999 间取值（或者一大一小取两个值进行计算），物种注释分类置信度在 0.9~0.99 间取值（或者一大一小取两个值进行计算，然后对结果进行综合）。

在提高参考数据库中各物种参考序列丰富度、选择合适的 OTU 聚类序列相似度和物种注释分类置信度的同时，增加目标数据的丰富度（比如在监测采样中进行有一定时空差异的重复采样），将可获得更全面的物种组成检出结果，并有望获得更准确的群落结构状况。对于鱼类种类组成的 eDNA 监测层面来讲，核心内容在于物种检出，无论是通过提升参考序列的丰富度覆盖尽可能多的种内变异，使 eDNA 所监测到的各物种的特定变异序列都能够获得可匹配的参考序列，还是通过增加 eDNA 所监测到的各物种的种内核苷酸序列变异种类，使 eDNA 所监测到的各物种的种内核苷酸序列变异中总有一个或者几个能够匹配到参考序列上，都能够实现物种检出的目的[3]。本研究中，多序列数据比单序列数据注释出的物种数多 5%~7%，注释出的 OTU 数多 78%~215%，物种注释正确占比高 4%左右（达 81.49%，图 3）。因此对于目标区域内的目标物种存在一定种内核苷酸序列变异，同时参考数据库里的参考序列对种内核苷酸序列变异的覆盖度不高的情况下，可以通过增加目标数据的丰富度（比如增加目标区域的时空差异性重复采样）以获得更全面的物种组成检出结果，并且还能获得更准确的群落结构组成状况。

基于宏条形码（meta-barcoding）的 eDNA 监测技术路径中，受制于物种注释算法，eDNA 序列物种注释可能存在并且长期存在注释错误和注释遗漏的问题。本研究对 2 个目标数据的分析注释结果显示，单序列数据和多序列数据中未被正确注释的鱼类物种均为本研究扩增获得但 nt 数据库中没有参考序列的物种（附表 2，附图 3b），说明在参考数据库中的参考序列和目标数据中的序列一致时，在物种注释过程中不存在某个物种被其它物种遮蔽的必然性。在 0.999、0.99 的 OTU 聚类序列相似度，0.99、0.9、0.8 的物种注释分类置信度（Blast 算法，序列一致性和序列覆盖度，两者取同一个值）水平，单个注释计算结果均有 15%以上的错误率，3%以上的遗漏率（附表 2，附图 3a，b），说明因为算法的内在局限性，单次计算存在一定的将一个物种的序列错误匹配到其它物种的参考序列上的概率。多序列数据在 0.999 的 OTU 聚类序列相似度和 0.99、0.9 的物种注释分类置信度下的物种正确注释比例为 81%左右，多序列数据在 0.99 的 OTU 聚类序列相似度和 0.9、0.8 的物种注释分类置信度下和单序列数据在 0.999 的 OTU 聚类序列相似度和 0.99、0.9 的物种注释分类置信度下的物种正确注释比例为 77%左右，单序列数据在 0.99 的 OTU 聚类序列相似度和 0.9、0.8 的物种注释分类置信度下的物种正确注释比例为 72%左右（附表 2，附图 3a，b），说明在单次计算中将一个物种的序列错误匹配到其它物种的参考序列上的概率还受目标数据、计算参数影响。

4 结论

针对鱼类的 eDNA 监测中，测序数据（线粒体 *COI* 基因的扩增子，引物为 mlCOIintF/jgHCO2198R，序列片段 320bp）的 OTU 聚类建议在 0.99~0.999 间取 OTU 聚类序列相似度值，对所获 OTU 用 Blast 算法进行参考 nt 库的物种注释，建议在 0.9~0.99 间取物种注释分类置信度（序列一致性和序列覆盖度，两者取同一个值），所获物种组成中可能仅有 70%~80%是确定正确的。为了提高注释结果的正确率，可以提高参考数据库中的物种覆盖度和各物种种内变异覆盖度（建立完善的本地参考数据库），增加目标区域的时空差异性重复采样进而增加目标数据的丰富度。但受制于物种注释算法，单次计算存在一定的将一个物种

的序列错误匹配到其它物种的参考序列上的概率，物种注释错误和注释遗漏的问题可能将长期存在。针对基于 *COI* 基因的鱼类 eDNA 监测中，在参考数据库、计算指标阈值、目标数据合适的情况下，物种注释错误的比例力求控制到 15%以内，物种注释遗漏的比例力求控制在 3%以内。随着参考数据库的完善、计算指标阈值的优化、目标数据的丰富，鱼类 eDNA 监测结果的物种组成有望更为准确地反映监测江段实际鱼类物种组成、各物种序列的相对丰度有望更为贴近监测江段实际鱼类群落结构。

参考文献

[1] Taberlet P, Coissac E, Hajibabaei M, *et al.*. Environmental DNA. *Molecular Ecology*. 2012, **21**(8): 1789-1793.

[2] Pawlowski J, Apothéloz-Perret-Gentil L, Altermatt F. Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*. 2020, **29**(22): 4258-4264.

[3] Hai Le Yang, Hui Zhang, Hao Du. A framework for standardizing the processes of eDNA monitoring and an accessible vision of the future. Journal of Lake Sciences. 2023, **35**(1): 12-31.[杨海乐，张辉，杜浩. eDNA监测方法标准化框架及未来图景. 湖泊科学. 2023, **35**(1): 12-31.]

[4] Deiner K, Bik H M, Mächler E, *et al.*. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*. 2017, **26**(21): 5872-5895.

[5] Coble A A, Flinders C A, Homyack J A, *et al.*. eDNA as a tool for identifying freshwater species in sustainable forestry: A critical review and potential future applications. *Science of the Total Environment*. 2019, **649**: 1157-1170.

[6] Deiner K, Yamanaka H, Bernatchez L. The future of biodiversity monitoring and conservation utilizing environmental DNA. *Environmental DNA*. 2021, **3**(3): 3-7.

[7] Dickie I A, Boyer S, Buckley H L, *et al.*. Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources*. 2018, **18**(5): 940-952.

[8] Nicholson A, Mcisaac D, Macdonald C, *et al.*. An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. *Environmental DNA*. 2020, **2**(3): 343-349.

[9] Itakura H, Wakiya R, Yamamoto S, *et al.*. Environmental DNA analysis reveals the spatial distribution, abundance, and biomass of Japanese eels at the river-basin scale. *Aquatic Conservation: Marine and Freshwater Ecosystems*. 2019, **29**(3): 361-373.

[10] Hänfling B, Lawson Handley L, Read D S, *et al.*. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*. 2016, **25**(13): 3101-3119.

[11] Nardi C F, Fernández D A, Vanella F A, *et al.*. The expansion of exotic Chinook salmon (Oncorhynchus tshawytscha) in the extreme south of Patagonia: an environmental DNA approach. *Biological invasions*. 2019, **21**(4): 1415-1425.

[12] Chen J, Chen Z, Liu S, *et al.*. Revealing an invasion risk of fish species in Qingdao underwater world by environmental DNA metabarcoding. *Journal of Ocean University of China*. 2021, **20**(1): 124-136.

[13] Ito G, Yamauchi H, Shigeyoshi M, *et al.*. Using eDNA metabarcoding to establish targets for freshwater fish composition following river restoration. *Global Ecology and Conservation*. 2023, **43**: e2448.

[14] Balasingham K D, Walter R P, Mandrak N E, *et al.*. Environmental DNA detection of rare and invasive fish species in two Great Lakes tributaries. *Molecular Ecology*. 2018, **27**(1): 112-127.

[15] Hai Le Yang, Lan Xin Xu, Qiong Zhou, *et al.*. Quantifying the spatial resolution of eDNA monitoring: a case study in Middle Yangtze River in mean-flow period. ChinaXiv. 2023(202303): 8735.[杨海乐，许兰馨，周琼等. eDNA监测空间分辨率量化的方法研究：以长江中游平水期为例. ChinaXiv. 2023(202303): 8735.]

[16] Hai Le Yang, Jin Ming Wu, Hui Zhang, *et al.*. Environmental DNA metabarcoding utilization efficiency in monitoring large river fish species composition: a case study in the Wuhan transect of the Yangtze River. Journal of Fishery Sciences of China. 2021, **28**(6): 796-807.[杨海乐，吴金明，张辉等. 大型河流中鱼类组成的eDNA监测效率:以长江武汉江段为例. 中国水产科学. 2021, **28**(6): 796-807.]

[17] Hai Le Yang, Li Shen, Yong Feng He, *et al.*. Status of aquatic organisms resources and their environments in Yangtze River system (2017-2021). Journal of Fisheries of China. 2023, **47**(2): 3-30.[杨海乐，沈丽，何勇凤等. 长江水生生物资源与环境本底状况调查（2017-2021）. 水产学报. 2023, **47**(2): 3-30.]

[18] Euclide P T, Lor Y, Spear M J, *et al.*. Environmental DNA metabarcoding as a tool for biodiversity assessment and monitoring: reconstructing established fish communities of north‐temperate lakes and rivers. *Diversity and Distributions*. 2021, **27**(10): 1966-1980.

[19] Meng Wang, Yi Yuan, Hai Yan Yu, *et al.*. Construction of barcode library of freshwater macroinvertebrate in China. Environmental Monitoring of China. 2022, **38**(1): 36-44.[王萌，苑艺，于海燕等. 中国淡水大型底栖无脊椎动物条形码数据库构建. 中国环境监测. 2022, **38**(1): 36-44.]

[20] Yang H L, Du H, Qi H F, *et al.*. Effectiveness assessment of using riverine water eDNA to simultaneously monitor the riverine and riparian biodiversity information. *Scientific Reports*. 2021, **11**(1): 24241.

[21] Ren Z, Wang F, Qu X, *et al.*. Taxonomic and functional differences between microbial communities in Qinghai Lake and its input streams. *Frontiers in Microbiology*. 2017, **8**(2319): 2319.